

Styx: Transactional Stateful Functions on Streaming Dataflows

Kyriakos Psarakis

George Siachamis

George Christodoulou

Marios Fragkoulis

Asterios Katsifodimos

Delft University of Technology

{initial.lastname}@tudelft.nl

ABSTRACT

Developing stateful cloud applications, such as high-throughput/low-latency workflows and microservices with strict consistency requirements, remains arduous for programmers.

The Stateful-Functions-as-a-Service (SFaaS) paradigm aims to serve these use cases. However, existing approaches either provide serializable transactional guarantees at the level of individual functions or separate application logic from the state and use inefficient transactional protocols. These design choices increase the execution latency, limiting the usability of SFaaS systems for stateful cloud applications.

In this paper, we present Styx, a novel SFaaS runtime that executes serializable transactions across functions with exactly-once guarantees. Styx is the first streaming dataflow-based runtime for SFaaS, offering application logic and state co-location, coarse-grained state persistence, and incremental checkpointing. Styx extends a deterministic transactional protocol to support an arbitrary call graph of stateful functions. It introduces a transaction-execution acknowledgment scheme that allows tracking a transactional workflow's SFaaS calls, guaranteeing atomicity and exactly-once processing. Experiments with the YCSB-T, TPC-C, and Deathstar benchmarks show that Styx outperforms state-of-the-art approaches by achieving at least one order of magnitude higher throughput while exhibiting near-linear scalability.

PVLDB Reference Format:

Kyriakos Psarakis, George Siachamis, George Christodoulou, Marios Fragkoulis, Asterios Katsifodimos. Styx: Transactional Stateful Functions on Streaming Dataflows. PVLDB, 14(1): XXX-XXX, 2020. doi:XX.XX/XXX.XX

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/delftdata/styx>.

1 INTRODUCTION

Despite the commercial offerings of the Functions-as-a-Service (FaaS) cloud service model, its suitability for low-latency stateful applications with strict consistency requirements, such as payment processing, reservation systems, inventory keeping, and low-latency business workflows, is quite limited. The reason behind this unsuitability is that current FaaS solutions are stateless, relying on external, fault-tolerant data stores (blob stores or databases)

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097. doi:XX.XX/XXX.XX

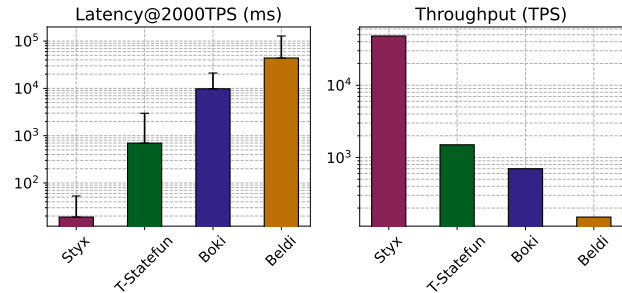


Figure 1: Styx outperforms the SotA by at least one order of magnitude in transactional workloads (§8). The figure shows median (bar)/99p (whisker) latency and throughput. For the latency plot, the input throughput is 2000 transactions per second (TPS), and for the throughput plot, we report the throughput that the systems achieve at 1-second latency.

for state management. In addition, while multiple frameworks can perform workflow execution (e.g., AWS Step Functions [49], Azure Logic Apps), primitives for *transactional* execution of such applications are non-existent, contrary to the interest of users [10, 32].

We argue that for FaaS offerings to become mainstream, they should include state management support for stateful functions according to the Stateful-Functions-as-a-Service (SFaaS) paradigm. Overall, a suitable offering for executing workflows of stateful functions would require *i*) low-latency and high-throughput execution, *ii*) end-to-end serializable transactional guarantees across multiple functions, and *iii*) high-level programming models, devoid of low-level primitives for locking and transaction coordination. To the best of our knowledge, no existing approach addresses all these requirements together.

The state-of-the-art transactional SFaaS with serializability guarantees Beldi [50], Boki [27], and T-Statefun [12], do support transactional end-to-end workflows but suffer from high commit latency, low throughput and inadequate programming models that leak transactional primitives to developers. The main reason behind their inefficiency is the separation of state storage and function logic, as well as the use of locking and Two-Phase Commit (2PC) [22] to coordinate and ensure atomicity of cross-function transactions.

This paper proposes Styx, a novel runtime for Stateful FaaS (SFaaS) that ensures exactly-once execution, low-latency, and high-throughput, while being able to execute arbitrary workflows of functions with end-to-end serializability guarantees. The guarantees offered by Styx enable the design of high-level programming models for cloud applications that obviate the need for failure management code associated with distributed systems and transaction processing inside application logic. In Styx, transactions are handled with a deterministic transaction protocol, avoiding costly 2PCs. As shown in Figure 1, Styx achieves at least one order of magnitude

lower median latency, two orders of magnitude lower 99p latency at 2000 transactions per second, and at least one order of magnitude higher throughput compared to the state of the art SFaaS systems [12, 27, 50] with end-to-end serializable guarantees.

In short, this paper makes the following contributions.

- Styx is a scalable, transactional, and fault-tolerant SFaaS runtime that leverages the streaming dataflow execution model and co-locates execution with state to offer exactly-once processing guarantees (§4).
- Styx provides multi-partition transactional serializability across arbitrary function calls (§5 and §6), extending the concept of deterministic databases for stateful functions and contributing a novel acknowledgment-sharing scheme (§5.3) to track function execution efficiently.
- Styx contributes a high-level programming model for transactional stateful functions, abstracting away transaction and failure management code (§3).
- Styx’s asynchronous incremental snapshots achieve fault-tolerance with minimal overhead to the system’s operation (§7), while Styx’s determinism enables early commit replies: the means to enable transactions to be reported as committed before a snapshot is committed to durable storage (§6.4).
- Styx outperforms the state-of-the-art [13, 27, 50] by at least one order of magnitude in all tested workloads in terms of throughput while achieving near linear scalability (§8).

Outline. The rest of the paper is organized as follows. §2 provides the background and highlights the motivation behind our work. Then, in §3, we present Styx’s programming model. In §4, we analyze the architecture of Styx. We delve into the sequencing & function execution workflow in §5, followed by §6, which discusses how we handle the deterministic commit of transactions. We end the technical description of Styx in §7 with the description of the fault tolerance mechanisms based on incremental snapshots. We devote §8 to our experimental methodology, results, and observations. Finally, we discuss the related work in §9 and conclude our paper in §10.

We make Styx publicly available: <https://github.com/delftdata/styx>

2 BACKGROUND AND MOTIVATION

Although relational database systems provide ACID transaction guarantees, many distributed real-world applications suffer from serious consistency issues when the responsibility of transaction execution is not ensured by a database system and is left to developers [10, 32].

In this section, we delve into the specifics of streaming dataflow systems design and argue that they can be extended to encapsulate the primitives required for executing stateful functions consistently and efficiently. We then show how the concept of deterministic databases can be extended for SFaaS, where transaction boundaries are unknown, unlike online transaction processing (OLTP). Finally, we argue that combining deterministic databases and dataflow systems can create a runtime that ensures atomicity, consistency, and scalability.

2.1 Dataflows for Stateful Functions

Stateful dataflows is the underlying execution model implemented by virtually all modern stream processors [8, 39, 41]. Besides being a great fit for parallel, data-intensive computations, stateful dataflows are the primary abstraction supporting workflow managers such as Apache Airflow [24], AWS Step Functions [49], and Azure’s Durable Functions [5]. We argue that stateful dataflows can make a suitable runtime for orchestrating general-purpose cloud applications for the following reasons.

Exactly-once Processing. Message-delivery guarantees are fundamentally hard to deal with in the general case, with the root of the problem being the well-known Byzantine Generals problem [34]. However, in the closed world of dataflow systems, exactly-once processing is possible [7, 8, 44]. As a matter of fact, the APIs of popular streaming dataflow systems, such as Apache Flink, require no error management code whatsoever (e.g., message retries or duplicate elimination with idempotency IDs).

Co-location of State and Function. The primary reason dataflow systems can sustain millions of events per second [8, 20] is that their state is partitioned across operators that operate on local state. While this co-location goes against the current economics of Cloud offerings that want state and computation to be decoupled, we believe that it is the primary vehicle to high-performance, and should be adopted by modern SFaaS runtimes.

2.2 Determinism & Transactions

Deterministic databases advance the level of transactional guarantees provided by traditional database systems. Given a set of database partitions and a set of transactions, a deterministic database [2, 46] will end up in the same final state despite node failures and possible concurrency issues.

Traditional database systems offer *serializable* guarantees, which allow multiple transactions to execute concurrently, ensuring that the database state will be equivalent to a serial execution of the transactions. Thus, deterministic databases guarantee not only serializability but also that a given set of transactions will have exactly the same effect on the database state despite transaction re-execution. This guarantee has multiple important implications [2] that, to the best of our knowledge, have not been leveraged by SFaaS systems thus far.

Deterministic Databases on Streaming Dataflows. Unlike 2PC, which requires rollbacks in case of failures, deterministic database protocols [37, 47] are “forward-only”: once the locking order [47] or read/write set [37] of a batch of transactions has been determined, the transactions are going to be executed and reflected on the database state, without the need to rollback changes. This is very much in line with how dataflow systems operate: events flow through the dataflow graph, from sources to sinks, without stalls for coordination. This match between deterministic databases and the dataflow execution model is the primary motivation behind Styx’s design choice to implement a deterministic transaction protocol on top of a dataflow system.

Non-Deterministic Functions on Streaming Dataflows. In principle, functions may encapsulate logic that makes the outcome of their execution non-deterministic. Examples of non-deterministic

operations are calls to external systems and the use of random number generators or time-related utilities. Styx, like state-of-the-art SFaaS systems [27, 50], currently supports functions with deterministic logic that always leads to the same output given the same input. That said, it is possible to embrace non-deterministic functions in Styx by tracking, recording, and persisting non-deterministic logic contained in them following the approach of Clonos [44].

2.3 Challenges & Open Problems

Despite their success and widespread applicability, dataflow systems need to undergo multiple changes before they can be used for transactional stateful functions. In the following, we list the challenges and open problems we tackle in this paper.

Programming Models. Dataflow systems at the moment are only programmable through functional-programming style dataflow APIs: a given cloud application has to be rewritten by programmers to match the event-driven dataflow paradigm. Although it is possible to rewrite many applications in this paradigm, it takes a considerable amount of programmer training and effort to do so. We argue that dataflow systems would benefit from object-oriented programming abstractions in order to be adopted by programmers for general cloud applications, such as microservices.

Support for Transactions. Although it has been recently shown that it is possible to introduce transaction coordination *on top of* a dataflow system [12], we argue that a more efficient implementation of transactions would require the dataflow system to be aware of transaction boundaries and to incorporate transaction processing into its state management and fault-tolerance protocols.

Deterministic OLTP and SFaaS. OLTP databases that use deterministic protocols like Calvin [47, 53] require each transaction’s read/write set a priori. Aria [37] extended deterministic transaction protocols by discovering the read-write sets of a transaction by first executing it. Aria works under the assumption that a transaction is encapsulated in a single-threaded function that can execute remote reads and writes from multiple partitions.

Arbitrary function calls enable programmers to take advantage of both the separation of concerns principle that is widely applied in microservice architectures [32], as well as code modularity. Although deterministic database systems have been proven to perform exceptionally well [2], the design and implementation of a deterministic transactional protocol for a SFaaS runtime with arbitrary function call graphs is non-trivial. More specifically, arbitrary function calls create complex call graphs that need to be tracked in order to establish a transaction’s boundaries before committing. At the same time, current deterministic transaction approaches employ a single thread that performs remote reads and writes instead of making arbitrary calls in parallel to functions that can mutate state.

Dataflows for Arbitrary-Workflow Execution. The prime use case for dataflow systems nowadays is streaming analytics. However, general cloud applications have different workload characteristics. Functions calling other functions and receiving return values introduce cycles in the dataflow graph. Such cycles can cause deadlocks and other flow-control issues that need to be dealt with [35].

```

1 from styx import Operator
2 from deathstar.operators import Hotel, Flight
3
4 Reservation = Operator('reservation', n_partitions=4)
5
6 @Reservation.register
7 async def make_reservation(context, flight_id, hotel_id, user_id):
8
9     context.call_async(operator=Hotel,
10                       function_name='reserve_hotel',
11                       key=hotel_id)
12     context.call_async(operator=Flight,
13                       function_name='reserve_flight',
14                       key=flight_id)
15
16     reservation = {"fid": flight_id, "hid": hotel_id, "uid": user_id}
17     await context.state.put(reservation)
18
19     return "Reservation Successful"

```

Figure 2: Deathstar’s[19] Hotel/Flight reservation in Styx. From lines 9-14, the `reserve_hotel` and `reserve_flight` functions are invoked asynchronously. Finally, in lines 16-17, the reservation information is stored. In Styx, the transactional and fault tolerance logic are handled internally.

In this work, we tackle these challenges and propose Styx: a novel dataflow system tailored to the needs of stateful functions, with built-in support for deterministic transactions and a high-level programming model.

3 PROGRAMMING MODEL

The programming model of Styx is based on Python and comprises operators that encapsulate partitioned mutable state and functions that operate on that. An example showcasing the programming model of Styx can be found in Figure 2.

3.1 Programming Model Notions

Stateful Entities. Similar to objects in object-oriented programming, entities in Styx are responsible for mutating their own state. Moreover, when a given entity needs to update the state of another entity, it can do so via a function call. Each entity bears a unique and immutable key, similar to Actor references in Akka [1], with the difference that entity keys are application-dependent and contain no information related to their physical location. The dataflow runtime engine (§4) uses that key to route function calls to the right operator that accommodates that specific entity.

Functions. functions can mutate the state of an entity. By convention, the context is the first parameter of each function call. Functions are allowed to call other functions directly, and Styx supports both synchronous and asynchronous function calls. For instance, in lines 9-11 of Figure 2, the instantiated reservation entity will call asynchronously the function ‘`reserve_hotel`’ of an entity with key ‘`hotel_id`’ attached to the Hotel operator. Similarly, one can make a synchronous call that blocks to wait for results. In this case, Styx will block execution until the call returns. Depending on the use case, a mix of synchronous and asynchronous calls can be used. Asynchronous function calls, however, allow for further optimizations that Styx applies whenever possible, as we describe in §5 and §6.

Operators. Each entity directly maps to a dataflow operator (also called a vertex) in the dataflow graph. When an *event* (i.e., function invocation) enters the dataflow graph, it reaches the operator holding the *function code* of the given entity as well as the *state* of that entity. In short, a dataflow operator can execute all functions of a given entity and store the state of that entity. Since operators can be partitioned across multiple cluster nodes, each partition stores a set of stateful entities indexed by their unique key. When a function of an entity is invoked (via an incoming event), the entity’s state is retrieved from the local operator state. Then, the function is executed using the arguments found in the incoming event that triggered the call.

State & Namespacing. As mentioned before, each entity has access only to its own state. In Styx, the state is *namespaced* with respect to the entity it belongs to. For instance, a given key "hotel153" within the operator `Hotel` is represented as: `entities://Hotel/hotel153`. This way, a reference to a given key of a state object is unique and can be determined at runtime when operators are partitioned across workers. Programmers can store or retrieve state through the context object by invoking `context.put()` or `get()` (e.g., in Line 17 of Figure 2). Styx’s context is similar to the context object used in other systems such as Flink Statefun, AWS Lambda, and Azure Durable Functions.

Transactions. A transaction in Styx begins with the invocation of an "entry" function. The subsequent functions that are called from the entry function (call graph) form a transaction that executes with serializable guarantees. Styx’s programming model allows transaction aborts by raising an uncaught exception. In the example of Figure 2, if a hotel entity does not have enough availability when calling the `'reserve_hotel'` function, the `'make_reservation'` transaction should be aborted, alongside potential state mutations that the `'reserve_flight'` has made to a flight entity. In that case, the programmer has to raise an exception as follows:

```

1 ...
2 # Check if there are enough rooms available in the hotel
3 if available_rooms <= 0:
4     raise NotEnoughSpace(f'No rooms in hotel: {context.key}')
5 ...

```

The exception is caught by Styx, which automatically triggers the abort/rollback sequence of the transaction where the exception occurred and sends the user-defined exception message as a reply.

Exactly-once Function Calling. A significant contribution of Styx is that it offers *exactly-once processing* guarantees: Styx reflects the state changes of a function call execution exactly-once. This way, programmers do not need to “pollute” their business logic with consistency checks, state rollbacks, timeouts, retries, and idempotency [31, 32]. We detail this capability in §7.

4 STYX’S ARCHITECTURE

In this section, we describe the components of Styx and the main design decisions we had to consider along the way as they appear in Figure 3.

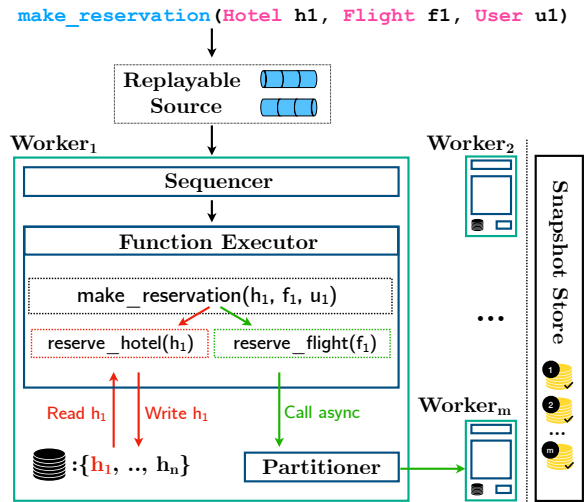


Figure 3: Stateful-Function execution in Styx.

4.1 Components

Coordinator. The coordinator manages and monitors the Styx workers. Its primary task is to manage the runtime state of the cluster (parallelism, dataflow state, partition locations, etc.) as well as perform scheduling and health monitoring. Currently, Styx implements round-robin scheduling, but we plan to add more complex schedulers in the future to be able to deal with load balancing. Styx monitors the cluster’s health using a typical heartbeat mechanism and initiates the fault-tolerance mechanism (§7) once a worker fails. Furthermore, it stores worker metrics that can be used by a dashboard to provide insights on the cluster, such as latency, the size of the sequences being processed, and partitioning imbalances.

Worker. As depicted in Figure 3, the worker is the primary component of Styx, managing networking, processing, and application state. At its core, it consists of two primary coroutines. The first coroutine ingests messages for its assigned partitions in a push-pull fashion from durable queue systems, such as Kafka, and sequences them. The second coroutine receives a sequence of transactions and initiates the transaction processing. By utilizing the coroutine execution model, Styx increases its efficiency since the most significant latency factor is waiting for network or state-access calls. Coroutines allow for single-threaded concurrent execution, switching between coroutines when one gets suspended during a network call, allowing others to make processing progress. Once the network call is completed, the suspended coroutine can resume processing.

Partitioning Stateful Entities Across Workers. Styx makes use of the entities’ key to distribute those entities and their associated state across a number of workers. By default, each worker is assigned a set of keys using hash partitioning.

Replayable Source. An integral part of the architecture to support the fault tolerance scheme that Styx employs is a replayable source. Other than an ingress to the system providing the function calls from external systems (e.g., from a REST gateway API), the source must be able to deterministically replay messages based on an offset when a failure occurs. This is necessary to produce the same

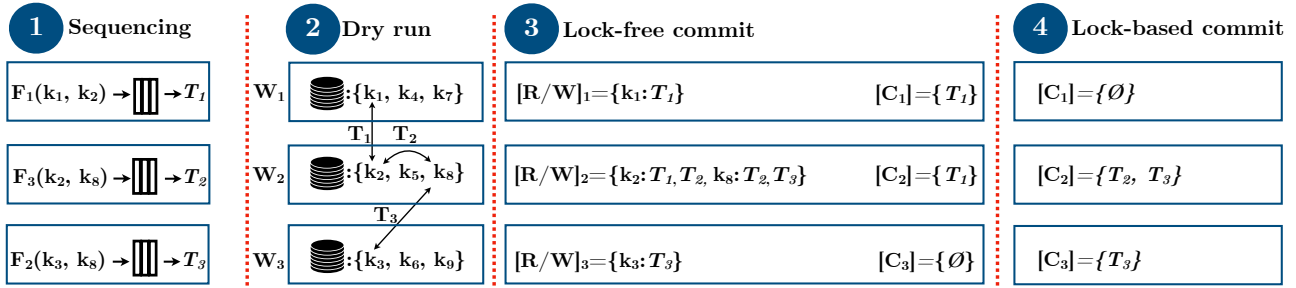


Figure 4: Transaction execution life-cycle in Styx. The pipeline is divided in 4 parts. First, each function is converted to a transaction and assigned a unique id by the sequencers. Afterwards, the transactions declare the keys they intend to interact with and determine their call graphs. As we can see, transactions conflict over accessing the same keys. Subsequently, we commit the transactions that do not participate in unresolved conflicts without locking. For example, we observe that $worker_1$ and $worker_2$ are capable to commit T_1 while it interacts with the same keys as T_2 ; although it has the lowest id. In the final part, we commit all the transactions by resolving the conflicts with a lock mechanism.

sequence of transactions after the recovery is complete and to enable early replies (§7). Currently, Styx uses Apache Kafka for that purpose.

Durable Snapshot Store. Alongside the replayable source, durable storage is necessary for storing the workers’ snapshots. Currently, Styx uses Minio, an open-source S3 clone, to store the incremental snapshots as binary data files.

4.2 Function/Transaction Execution Pipeline

In Styx, the terms function and transaction can be used interchangeably. A function triggers a call graph of function invocations within the system. Each such function can have its own state; the side-effects of the call graph of function invocations are transactionally committed to the state. In Figure 3, once a `make_reservation` function enters the system, it is persisted and replicated in the replayable source. Then, based on the partitioning, a worker ingests the function call into its sequencer and transforms it into a transaction by assigning a TID. The next step is to execute the transaction based on the transactional protocol detailed in §5 and §6. In the `make_reservation` case, the transaction consists of two functions: `reserve_hotel` and `reserve_flight`. For this example, let us assume that `reserve_hotel` is a local function call and `reserve_flight` a remote. `reserve_hotel` will execute locally (in an asynchronous fashion using co-routines) and apply state changes. In contrast, `reserve_flight` will execute asynchronously on a remote worker, applying state changes on the remote with zero state transfer other than the function parameters.

5 SEQUENCING & FUNCTION EXECUTION

The deterministic execution of functions with serializable guarantees requires a sequencing step that assigns a sequence ID (TID), which, in combination with the read/write (RW) sets, can be used for conflict resolution (§6). The challenge that we tackle in this section is determining the boundaries of transactions (i.e., when transactional function executions begin and complete), which emerges from the execution of arbitrary function call graphs §5.3.

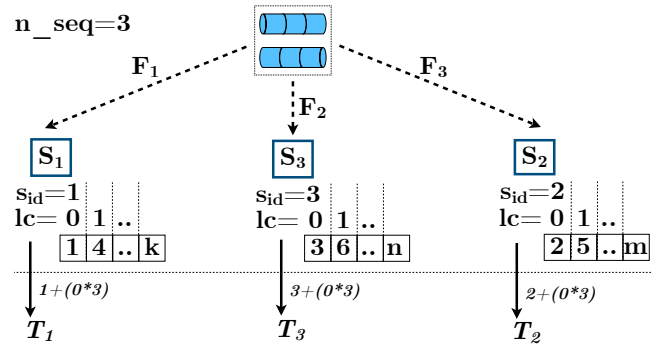


Figure 5: Example of TID assignment in Styx in the case of three sequencers. Their identifiers $\{1, 2, 3\}$ lead to the following sequences: $S_1 = \{1, 4, \dots, k\}$, $S_2 = \{2, 5, \dots, m\}$, $S_3 = \{3, 6, \dots, n\}$ following the formula expressed in Equation (1)

5.1 Function Sequencing

In this section, we discuss the sequencing mechanism (1) of Styx. Deterministic databases ensure the serializable execution of transactions by creating a global sequence. In Calvin [47], the authors propose a partitioned sequencer that retrieves the global sequence by communicating across all partitions, performing a deterministic round-robin.

Eliminating Sequencer Synchronization. Instead of the original sequencer of Calvin that sends $O(n^2)$ messages for the deterministic round-robin, Styx adopts a method similar to Mencius [38], allowing Styx to acquire a global sequence without any communication between the sequencers ($O(1)$). This is achieved by having each sequencer assign unique transaction identifiers (TIDs) as follows:

$$TID_{sid,lc} = sid + (lc * n_seq) \quad (1)$$

where $sid \in \mathbb{N}_1$ is the sequencer id assigned by the Styx coordinator in the registration phase, $lc \in \mathbb{N}_0$ is a local counter of each sequencer specifying how many TIDs it has assigned thus far and $n_seq \in \mathbb{N}_1$ is the total number of sequencers in the Styx cluster. As an example in Figure 5 there are three sequencers with $sids = \{1, 2, 3\}$ leading to the following sequences: $S_1 = \{1, 4, \dots, k\}$, $S_2 = \{2, 5, \dots, m\}$,

$S_3 = \{3, 6, \dots, n\}$. Thus, this arrangement leads to the same result as the deterministic round-robin schedule without any messaging between the sequencers.

In the example of Figure 4, the sequencers of the three workers will sequence $F_1(k_1, k_2)$, $F_2(k_3, k_8)$ and $F_3(k_2, k_8)$ to T_1 , T_3 and T_2 respectively.

Sequence Imbalance. A consideration with the approach used in Styx is the possibility of imbalance between the produced sequences. In the worst-case scenario where the total traffic in an epoch is directed to a single sequencer, the local counter lc will increment by that amount. Then, in the next epoch, that sequencer will produce larger TIDs than the other sequencers, and consequently, the new transactions it receives will have a lower priority for execution. In concurrency-related conflicts, this will increase latencies for that worker node. To avoid this issue, we pass all lc within the commit message so that each worker can set its local counter (lc) as the max of all, leading to balanced sequences in every epoch.

Replication and Logging. There is no need for replication and logging of the sequence within Styx since the input is logged and replicated within the replayable source. After replaying in case of failure, the sequencers will produce the exact same sequence.

5.2 Call-Graph Discovery: Dry Run

After sequencing, Styx needs to execute the sequenced transactions and determine their call graphs and read/write sets (2). To this end, the function execution runtime ingests a given sequence of transactions to process in a given epoch. The number of transactions per epoch is either set by a timeout (by default, Styx ingests all transactions that have arrived every millisecond) or by a configurable maximum number of transactions that can run per epoch (by default 1000 transactions per epoch). We have chosen an epoch-based approach since processing the incoming transactions in a batch will lead to higher throughput.

Styx’s runtime executes all the sequenced transactions on a snapshot of the data to retrieve the read/write sets. Transactions that span multiple workers will implicitly change the read/write sets of the remote workers via function calls. There is an additional issue related to discovering the read/write set of a transaction: before the functions execute, the call graph of the transaction and the remote function calls are unknown. This is an issue because the protocol requires all transactions to be completed before proceeding to the next phase. To tackle this problem, Styx proposes a function chaining protocol that is explained in more detail in §5.3.

After this phase, all the stateful functions that comprise transactions would have finished execution and the read/write sets will be known. In Figure 4, transactions T_1 , T_2 , and T_3 will execute and create the following RW sets: $Worker_1 \rightarrow \{k_1 : T_1\}$, $Worker_2 \rightarrow \{k_2 : T_1, T_2 \text{ and } k_8 : T_2, T_3\}$ and $Worker_3 \rightarrow \{k_3 : T_3\}$.

5.3 Function Acknowledgement Shares

In the SFaaS paradigm, the call graph that forms a transaction is unknown; functions could be coded by different people, calling other functions written by others, and so on. This makes it difficult to determine when a transaction has finished processing and is necessary because phase 3 can only start after all transactions have

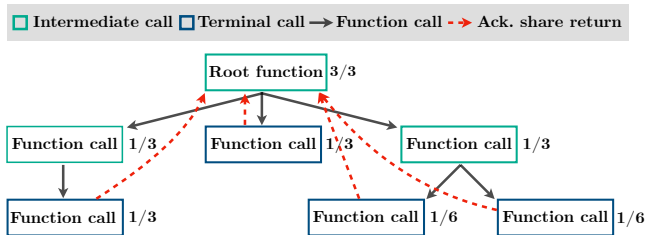


Figure 6: Asynchronous function call chains. A given root function call may invoke other functions throughout its execution. The original acknowledgment (3/3) splits into parts as the function execution proceeds, and each function receives its own ack-share. For instance, in this function execution, the root function calls three other functions, thus splitting the ack-share into three equal parts. The same applies to subsequent calls, where the caller functions further split their ack-share. The sum of ack-shares of terminal (blue) calls (i.e., function calls that do not perform further calls) adds to exactly 3/3, which allows the root function to report the completion of execution.

finished processing. To that end, each asynchronous function call of a given transaction is assigned an `ack_share`. A given function knows how many shares to create by counting the number of asynchronous function calls during its runtime. The caller function then sends the respective acknowledgment shares to the downstream functions. For instance, in Figure 6, the transaction entry-point (root of the tree) calls three remote functions, splitting the `ack_share` into three parts ($3 \times 1/3$). The left-most function invokes only one other function and passes to it its complete `ack_share` (1/3). The middle function does not call any functions, so it returns the share to the transaction entry point (root), and the right-most function calls two other functions, splitting its share (1/3) to $2 \times 1/6$. After all the function calls are complete, the entry-point should have collected all the shares. When the sum of the received shares adds to 1, the root/entry-point function can safely deduce that execution is complete.

This design was devised for two reasons: i) if every participating function just sent an ack when it was done, the root would not know how many acks to expect in order to decide whether the entire execution is finished, and ii) if we used floats instead of fractions we could stumble upon a challenge related to adding floating point numbers. For instance, if we consider floating point numbers in the aforementioned example of the three function calls, the sum of all shares would not equal 1, but 0.99 since each share contributes 0.33. Hence, we cannot accurately round inexact division numbers, and therefore, Styx makes use of fraction mathematics instead.

A solution close to the `ack_share` is the one of distributed futures [48]. However, it would not work in the SFaaS context because it either requires information about the entire call graph for it to work asynchronously, or it would need to create a chain of futures that would make it synchronous and hence it would introduce high latency for our use case.

6 COMMITTING TRANSACTIONS

After the completion of an epoch’s dry run, Styx needs to determine which transactions will commit and which will abort based on the transactions’ RW sets and TIDs. To this end, this section presents the two commit phases: an optimistic lock-free phase and a lock-based phase that leverages the RW sets to execute a commit phase by acquiring locks ordered by TID. To make the second phase faster, we have devised a call-graph caching scheme that can reuse the dry run’s call graph to avoid re-executing long call-graph chains (§6.3).

6.1 Lock-free Commit Phase

In case of conflict (i.e., a transaction t writes a key that another transaction t' also reads or writes on) similarly to [37], only the transaction with the lowest transaction ID will succeed to commit (3). The transactions that have not been committed are put in a queue to be executed in the next phase (4) (maintaining their previously assigned ID). In addition, workers broadcast their local conflicts to every other worker: this way, every worker retains a global view of all the aborted/rescheduled transactions and can decide, locally, which transactions can be committed. Finally, note that transactions can also abort, not because of conflicts, but due to application logic causes (e.g., by throwing an exception). In that case, Styx removes the related entries from the read/write sets to reduce possible conflicts further.

In this phase, all the transactions that have not been part of a conflict apply their writes to the state and reply back to the clients. In the example shown in Figure 4, only T_1 can commit in W_1 and W_2 due to conflicts in the RW sets of W_2 regarding T_2 and T_3 ; more specifically in keys k_2 and k_8 .

6.2 Lock-based Commit Phase

In the previous phase, (3), only transactions without conflicts can be committed. We now explain how Styx deals with transactions that have not been committed in a given epoch due to conflicts (4). First, Styx trivially acquires locks in a given sequence ordered by transaction ID. Then, it reruns all transactions concurrently since all the read/write sets are known and commits them. However, if a transaction’s read/write set changes in this phase, Styx aborts the transaction and recomputes its read/write set in the next epoch. Now, in Figure 4, W_2 can sequentially acquire locks for T_2 and T_3 , leading to their commits in W_2 and W_3 . In §6.3, we present an optimization for this phase that we introduced in Styx, which reduces its latency impact even further.

6.3 Call-Graph Caching

As depicted in Figure 4, the lock-based commit phase (4) is used to execute any transactions that did not commit during the lock-free commit phase (3). Thus, by the time the lock-based commit phase is ready to execute, the state of the database may have changed since the lock-free commit. As a result, function invocations need to be re-executed to take into account the data updates.

On the left of Figure 7, we depict such a function invocation. At time t_0 , F_1 is invoked, which in turn invokes two function chains: $F_1 \rightarrow F_2 \rightarrow F_4 \rightarrow F_6$ and $F_1 \rightarrow F_3 \rightarrow F_5$. Once the two function chains finish their execution (on time t_3 and t_2 respectively), they can acknowledge their termination to the root call F_1 .

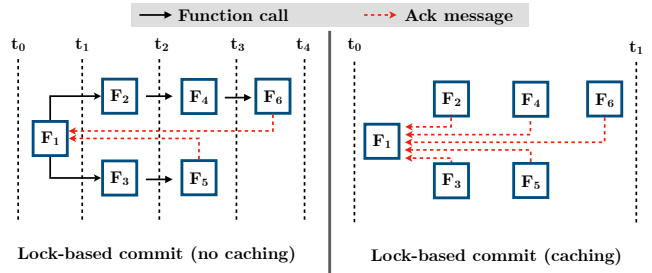


Figure 7: If no function caching is performed (left), the transaction execution will execute a deep call graph, and the messages sent will be equal to the number of function calls (5) in addition to the acks (2). Styx’s function caching optimization (right) will lead to a concurrent function execution in the lock-based commit phase, between t_0 and t_1 , and send only five acks.

Potential for Caching. During our early experiments, we noticed cases where F_1 is invoked and the parameters with which it calls F_2 (and in turn the invocations across the $F_1 \rightarrow \dots \rightarrow F_6$ call chain) do not change. The same applied to the RW set of those function invocations; the RW sets remained unchanged. Since Styx tracks those call parameters as well as the functions’ RW sets, Styx can simply cache input parameters during the lock-free commit phase and re-use them during the lock-based commit in order to avoid long sequential re-executions along the call chains. This case is depicted on the right of Figure 7: the function-call chain does not need to be invoked in a sequential manner from F_1 all the way to F_6 , leading to high-latency. Instead, the individual workers can simply re-invoke those function calls locally and in parallel. As a result, all functions can execute in parallel and save on latency and network overhead ($t_4 - t_1$ in Figure 7).

Conditions for Parallel Function Re-invocation. Intuitively, if the parameters with which e.g., F_2 is called, and the RW set of F_2 remains the same, we can safely assume that function F_2 can be invoked in parallel without having to be invoked sequentially by F_1 . If those functions are successfully completed and acknowledge their completion to the root function F_1 , it means that the transaction can be committed. To the contrary, if the RW set of any of the functions $F_1 - F_6$ changes, or the parameters of any of the functions along the call chains change, the transaction will need to be fully re-executed. Styx in that case will have to re-schedule that transaction in the next epoch.

6.4 Early Commit Replies via Determinism

Implementing Styx as a fully deterministic dataflow system offers a set of advantages involving the ability to communicate transaction commits to external systems (e.g., the client), even before the state snapshots are persisted to durable storage. A traditional transactional system can respond to the client only when *i*) the requested transaction has been committed to a persistent, durable state or *ii*) the write-ahead log is flushed and replicated. In Styx’s case, that would mean when an asynchronous snapshot completes (i.e., is persisted to stable storage such as S3), leading to high latency.

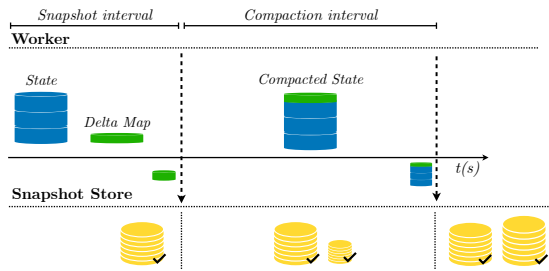


Figure 8: Incremental snapshots with Delta Maps in Styx.

Since Styx implements a deterministic transactional protocol executing an agreed-upon sequence of transactions among the workers, after a failure, the system would run the same transactions with exactly the same effects. This determinism enables Styx to give early commit replies: the client can receive the reply even before a persistent snapshot is stored. The assumption here is that the message broker that persists the client requests per partition will provide to the Styx sequencers the transactions in the same order after replay as under the system’s normal operation prior to the failure, a guarantee that is typically provided by most modern message brokers including Apache Kafka that Styx currently uses.

7 FAULT TOLERANCE

To ensure atomicity at the level of workflow execution, existing SFaaS systems perform fine-grained fault tolerance [27, 50]: each function execution is logged and persisted in a workflow log before the next function is called. This requires a round-trip to the logging mechanism for each function call, which adds significant latency to function execution. Styx opts for a coarse-grained fault tolerance mechanism. Instead of logging each function execution, it adopts a variant of existing checkpointing mechanisms used in streaming dataflow systems [7, 9, 44]: Styx asynchronously snapshots state, enabling very low-latency function execution. We describe this fault tolerance mechanism below.

7.1 Incremental Snapshots & Recovery in Styx

The snapshotting mechanism of Styx resembles the approach of many streaming systems [3, 7, 20, 26], that extend the seminal Chandy-Lamport snapshots [9]. Modern stream processing systems checkpoint their state by receiving snapshot barriers at regular time intervals (epochs) decided by the coordinator. In contrast, Styx leverages an important observation: workers do not need to wait for a barrier to enter the system in order to take a snapshot since the natural barrier in a transactional epoch-based system like Styx is at the end of a transaction epoch.

To this end, instead of taking snapshots periodically by propagating markers across the system’s operators, Styx aligns snapshots with the completion of transaction epochs to take a consistent cut of the system’s distributed state – including the state of the latest committed transactions, the offsets of the message broker and, the sequencer counter (lc). When the snapshot interval triggers, Styx makes a copy of the current state changes to a parallel thread and persists incremental snapshots asynchronously, allowing Styx to continue processing incoming transactions while the snapshot

operation is performed in the background. In combination with a replayable source for the incoming transactions (e.g., a Kafka message queue), this approach guarantees the processing of transactions exactly-once, even under failures. In case of a system failure, Styx *i*) rolls back to the epoch of the latest complete snapshot, *ii*) loads the snapshotted state, *iii*) rolls back the replayable source’s topic partitions (that are aligned with the Styx operator partitions) to the offsets at the time of the snapshot, *iv*) loads the sequencer counter, and finally, *v*) verifies that the cluster is healthy before executing a new transaction epoch.

Incremental Snapshots & Compaction. Each snapshot stores a collection of state changes in the form of *delta maps*. A delta map is essentially a hash table that tracks the changes in a worker’s state in a given epoch. When a snapshot is taken, only the delta map containing the state changes of the current epoch is snapshotted. To avoid having to track changes across delta maps, Styx performs periodical compactions where the deltas are merged in the background, as shown in Figure 8. In the worst case, Styx will have to merge one delta map to recover a consistent snapshot in case of a failure.

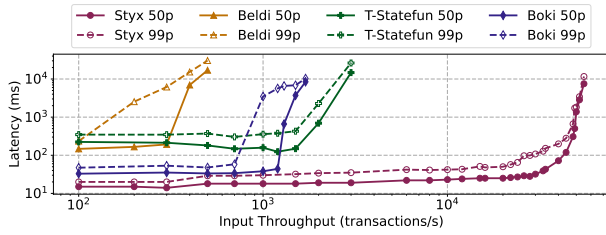
7.2 Exactly-once Output Guarantees

Another common challenge in the fault tolerance of streaming systems is the one of exactly-once output [16] in the presence of failures, which is hard to solve for low-latency use cases. For example, in Apache Flink’s [8] exactly-once output configuration, clients can only retrieve responses after those are persisted in a transactional sink. This arrangement is sufficient for streaming analytics but not for low-latency transactional workloads as discussed previously in §6.4.

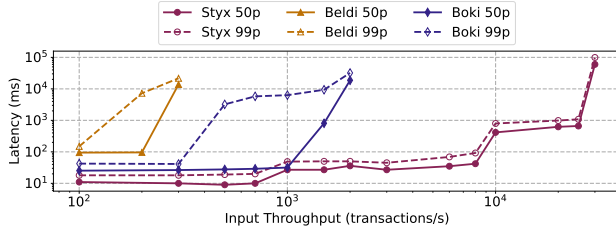
To solve that, Styx persists the output offsets on the output queue, and during recovery Styx *i*) reads the last offset of the message broker, *ii*) compares it with the offset of the message broker persisted in the snapshot to calculate the batch of transactions for processing, *iii*) retrieves the TIDs attached in the replies of the output queue, and *iv*) deduplicates the outputs that have already been sent to the output queue, before it continues processing. Note that this deduplication strategy is based on the fact that TIDs have been assigned deterministically.

7.3 Replication and High Availability

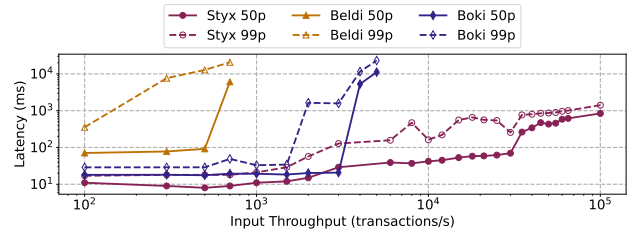
Although Styx achieves fault tolerance by storing snapshots in a replicated fault-tolerant blob store (e.g., Minio / S3), it does not yet support other forms of active replication for high availability scenarios. However, Styx in the future can easily inherit the replication mechanisms of deterministic databases. The design of deterministic transaction protocols, such as Calvin [47], feature state replicas that require no explicit synchronization. First, the sequencer replicas need to agree on the order of execution. This is a relatively cheap synchronization step (using Paxos [33] or Raft [42]), as it does not require the exchange of state, only transaction IDs. After that, the deterministic sequencing algorithm guarantees that the resulting state will be the same across partition/worker replicas by virtue of all replicas executing state updates in the same order.



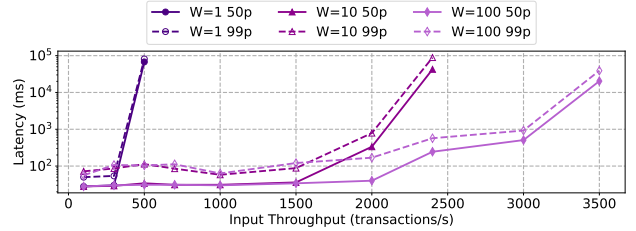
(a) YCSB-T (uniform).



(c) Deathstar Movie Review.



(b) Deathstar Travel Reservation.



(d) TPC-C on Styx with 1, 10, and 100 warehouses.

Figure 9: Evaluation in different scenarios. T-Statefun does not support range queries required by the Deathstar workloads. TPC-C is only supported by Styx

8 EVALUATION

In this section, we evaluate Styx by answering the following questions:

- How does Styx compare to State-of-the-Art serializable transactional SFaaS systems?
- How does Styx perform under skewed workload?
- How well does Styx scale?
- How much does the snapshotting mechanism affect performance?

8.1 Setup

Systems Under Test. In the evaluation, we only consider SFaaS systems that provide serializable transactional guarantees. At the time of writing, those are: *Beldi* [50], *Boki* [27] and *T-Statefun* [13]. *Beldi/Boki*. Both systems use Nightcore [28] as their function runtime, and both use a variant of two-phase commit and store the data in an AWS-managed DynamoDB with enabled autoscaling. *T-Statefun*. maintains the state and the coordination of the two-phase commit protocol within an Apache Flink cluster and sends the relevant state to remote stateless function runtimes for execution. For fault tolerance, it relies on a RocksDB state backend that performs incremental snapshots.

Styx. Styx core is implemented in Python 3.12 and uses coroutines to enable asynchronous concurrent execution. Styx uses ZeroMQ¹ for messaging between the workers. The system’s ingress point is another message queue, currently Apache Kafka, with important capabilities. It needs to support replayable partitioned fault-tolerant streams that can align with the partitioning of Styx’s operators.

¹<https://zeromq.org/>

Scenario	#keys	Function Calls	Transactions %
YCSB-T	10k	2	100%
Deathstar Movie	2k	7	0%
Deathstar Travel	2k	3	0.5%
TPC-C	1m-100m	8 / 20-50	100%

Table 1: Workload characteristics.

Finally, we use Minio²/S3 as a remote persistent store for the incremental snapshots. Unless stated otherwise, the snapshot fault-tolerance mechanism is enabled (10-second intervals) across all experiments shown in this section.

Workloads/Benchmarks. Table 1 summarizes the three workloads used in the experiments.

YCSB-T [15]. We use a variant of YCSB-T [15] from T-Statefun’s [12] evaluation. Each transaction consists of two reads and two writes. The concrete scenario is the following: first, we create 10,000 bank accounts (keys) and perform transactions in which a debtor attempts to transfer credit to a creditor. This transfer is subject to a check on whether the debtor has sufficient credit to fulfill the payment. If not, a rollback needs to be performed. The selection of a relatively small number of keys is deliberate: we want to assess the systems’ ability to sustain transactions under high contention. In addition, for the experiment depicted in Figure 10 (skewed distribution), we select the debtor key based on a uniform distribution and the creditor based on a Zipfian distribution, where we can vary the level of contention by modifying the Zipfian coefficient.

Deathstar [19]. We employ Deathstar [19], as adapted to SFaaS workloads by the authors of Beldi [50]. It consists of two workloads: *i*) the Movie workload implements a movie review service where users write reviews about movies. *ii*) the Travel workload implements a travel reservation service where users search for hotels

²<https://min.io/>

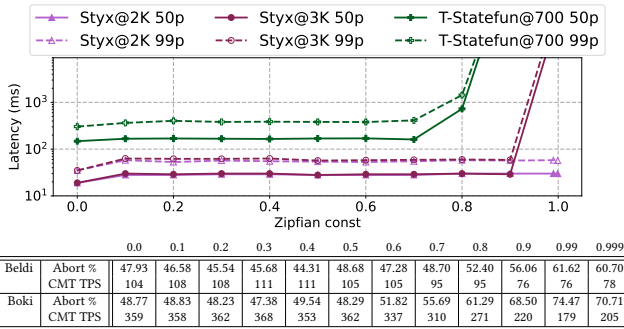


Figure 10: Latency evaluation for varying levels of contention with YCSB-T (skewed). We run Styx twice to show its behavior under contention clearly. Beldi and Boki execute at 200 and 700 TPS input rates, respectively. Abort percentage due to no-wait die protocol and effective output throughput (i.e., Committed TPS) during the Zipfian workload.

and flights, sort them by price/distance/rate, find recommendations, and transactionally reserve hotel rooms and flights. Both Deathstar workloads follow a uniform distribution. Note that T-Statefun could not run in this set of experiments since it does not support range queries.

TPC-C [36]. A transactional benchmark that targets OLTP systems is TPC-C [36]. To use it in our evaluation, we had to adapt SQL queries into the SFaaS paradigm, splitting the NewOrder transaction into 20-50 function calls (one call for each item in the NewOrder transaction) and the Payment transaction into eight function calls. TPC-C scales in size/partitions by increasing the number of warehouses represented in the benchmark. While a single warehouse represents a skewed workload (all transactions will hit the same warehouse), increasing the number of warehouses decreases the contention allowing for higher throughput and lower latency. Note that the TPC-C experiments do not include Beldi, Boki or T-Statefun: despite our efforts, the existing systems could not handle the complexity of the TPC-C workload (complex call graph) in order to produce meaningful experimental results for the needs of this paper.

Resources. For Beldi/Boki, T-Statefun and Styx, we assigned a total of 112 CPUs with 2GBs of RAM per CPU, matching what is presented in the original Boki paper [27]. Unless stated otherwise, Styx and T-Statefun are configured to perform incremental snapshots every 10 seconds. All external systems i.e., DynamoDB (Beldi, Boki), Minio, and Kafka (transaction ingress for Styx, T-Statefun) are configured with three replicas for fault tolerance.

External Systems. Boki and Beldi make use of a fully managed and autoscaled DynamoDB instance (configured to use 3 replicas) at AWS. The resources occupied by DynamoDB are additional to the 112 CPUs assigned to Boki and Beldi. Similarly, the resources assigned to Styx and T-Statefun do not include their snapshot store, Minio/S3.

Metrics. We are mainly interested in two important metrics.

Input throughput. represents the number of transactions that are submitted, per second, to the system under test. As the input

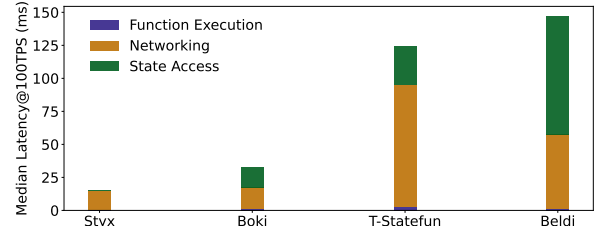


Figure 11: Performance breakdown of all systems.

throughput increases, throughout an experiment, we expect the latency of individual transactions to increase, until aborts start to manifest due to contention or high load.

Latency. represents the time interval between submitting a transaction, and the reported time when the transaction committed/aborted. In Styx and T-Statefun the latency timer starts when a transaction is submitted in the input queue (Kafka). The end of the transaction is when the system reports the transaction as committed/aborted in the output queue. Similarly, in Beldi and Boki the latency is the time since the input gateway has received a transaction, and the time that the gateway reports that the transaction has been committed/aborted.

8.2 Latency vs. Throughput

We first study the latency-throughput tradeoff of all systems. We retain the resources given to the systems constant (112 CPUs) while we progressively increase the input throughput. We monitor the transaction latency. As depicted in Figure 9, Styx outperforms its competitors by at least an order of magnitude. Specifically, Styx achieves a performance improvement of $\sim 20x$ against T-Statefun, which ranks second, in YCSB-T (Figure 9a) in terms of throughput, $\sim 30x$ against Boki in Deathstar’s travel reservation workload, (Figure 9b) and $\sim 35x$ against Boki in Deathstar’s movie review Figure 9c) workload. Finally, in the TPC-C benchmark (Figure 9d), which is the most demanding of all, as we increase the input throughput for different number of warehouses, we observe that Styx can handle up to 3K TPS with sub-second 99th percentile latency (100 warehouses).

Styx and T-Statefun executed without any concurrency-related aborts. On the contrary, Beldi and Boki utilize a no-wait-die concurrency control mechanism, which leads to a significant amount of aborts as the throughput increases, making the latency measurements incomparable. On this ground, in Figure 10 we plot the results of Styx and T-Statefun and present the performance of Beldi and Boki including their abort rates in a separate table.

We observe the following: *i*) at the highest level of contention (Zipfian at 0.999) Styx achieves 2000 TPS, outperforming the rest by $\sim 5-10x$ in terms of effective throughput (see table in Figure 10), *ii*) both Beldi and Boki abort more transactions as the level of contention increases ($\sim 40-70\%$), which significantly impacts their effectiveness, and *iii*) Styx manifests an increase in latency only in high levels of contention (Zipfian > 0.99) while executing at $\sim 4x$ higher throughput than the rest.

Systems’ Performance Breakdown. In Figure 11, we show where the systems under test spend their processing time. We use YCSB-T

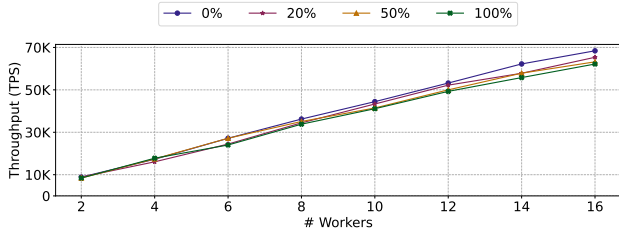


Figure 12: Scalability of Styx on YCSB-T with varying percentages of multi-partition transactions.

for this purpose since it is the benchmark with the lowest function execution complexity, making it suitable for testing how networking and state access affect performance. We measured the median latency while all the systems were running at 100 TPS for 60 seconds, and averaged the proportions of function execution, networking, and state access across all committed transactions. The key observations are: *i*) Styx’s co-location of processing and state led to near-zero state access latency and; *ii*) Styx’s asynchronous networking allows for very low network latency.

Experiments Takeaway. The rather large performance gains of Styx across all experiments are enabled by the following four properties and design choices: *i*) the co-location of processing and state with efficient networking as shown in Figure 11 contrary to the other systems that have to transfer the state to their function execution engines; *ii*) the asynchronous snapshots with delta maps for fault tolerance compared to the replication of Beldi/Boki and the LSM-tree-based incremental snapshots of T-Statefun; *iii*) the efficient transaction execution protocol employed in Styx compared to the two-phase commit used by Beldi, Boki, and T-Statefun; and *iv*) the underlying performance of Styx’s streaming dataflow engine.

8.3 Scalability

In this experiment, we tested the scalability of Styx by increasing the number of Styx workers (1 CPU each). We measure the maximum throughput on YCSB-T. The goal is to calculate the speedup of operations as the input throughput and number of workers scale together. In addition, we set a parameter to control the percentage of multi-partition transactions in the workload. In Figure 12, we observe that in all settings, Styx retains near-linear scalability. The percentage of multi-partition transactions in the workload does not affect performance significantly, albeit YCSB-T consists of two key accesses per transaction.

8.4 Snapshotting Microbenchmarks

Effect of Snapshots. In Figure 13 we depict the impact of the asynchronous incremental snapshots to Styx’s performance. In both figures, we display when a snapshot starts and ends. The state includes 1 million keys with a 1-second snapshot interval. Styx is deployed with four 1-CPU workers, and the input transaction arrival rate is fixed to 3K YCSB-T TPS. In Figure 13a, we observe that during a snapshot operation, Styx shows no performance degradation in the output throughput as it closely follows the input. In Figure 13b, we observe a minor increase in the end-to-end latency

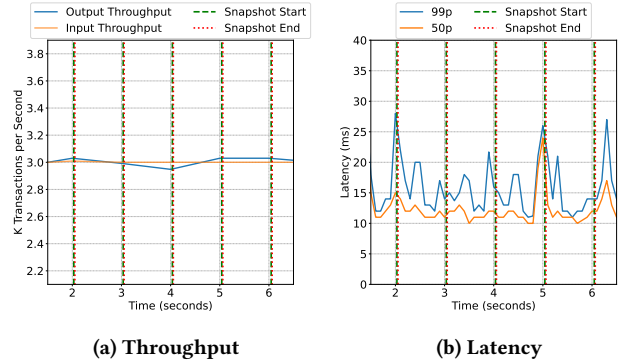


Figure 13: Styx’s snapshotting impact on performance at 1-second intervals.

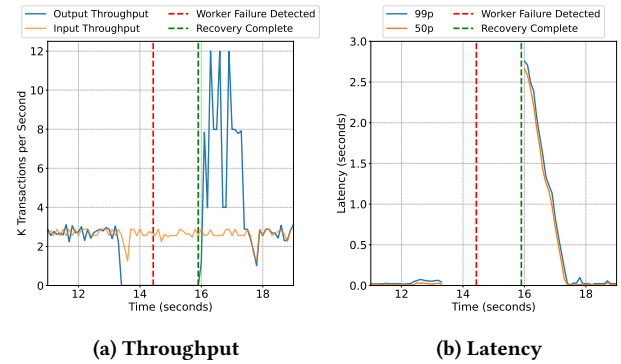


Figure 14: Styx’s behavior during recovery.

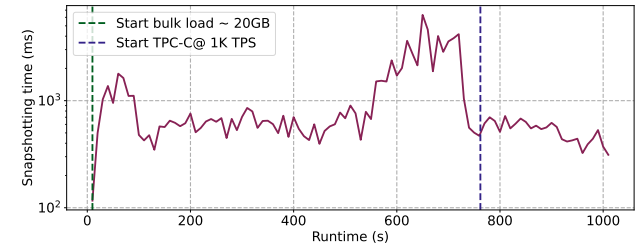


Figure 15: Behaviour of incremental snapshots on Styx with ~20GB TPC-C state.

in some snapshots. This is normal, as the concurrent snapshotting thread competes with the transaction execution thread during snapshotting, and it also has to block the transaction execution thread momentarily in order to copy the corresponding operator’s delta state.

Recovery Time. In Figure 14, we evaluate the recovery process of Styx with the same parameters as in Figure 13. We reboot a Styx worker at ~13.5 seconds. It takes Styx’s coordinator roughly a second to detect the failure. Then, after the reboot, the coordinator re-registers the worker and notifies all workers to load the last complete snapshot, merge any uncompact deltas, and use the message broker offsets of that snapshot. The recovery time is also

observed in the latency (Figure 14b) that is ~2.5 seconds (time to detect the failure + time to complete recovery). In terms of throughput (Figure 14a), we observe Styx working on its maximum throughput after recovery completes to keep up with the backlog and the input throughput.

Effect of Large State Snapshots. In Figure 15, we test the snapshotting mechanism against a larger state of 20 GB from TPC-C using a bigger Styx deployment of 100 1-CPU workers at 10-second checkpoint intervals. From 0 to the 750-second mark, Styx is importing the dataset. Since there are no small deltas (importing is an append-only operation), snapshotting is more expensive than the normal workload execution, where only the deltas are stored in the snapshots. The rise in latency at ~550 seconds corresponds to the loading of the largest tables (Stock and Order-Line) in the system. After loading the data and starting the transactional workload at 1000 TPS, we observe a drop in latency due to fewer state changes within the delta maps.

9 RELATED WORK

Transactional SFaaS. SFaaS has received considerable research attention and open-source work. Transactional support alongside fault tolerance guarantees (that popularized DBMS systems) is necessary to widen the adoption of SFaaS. Existing systems fall into two categories: i) those that focus on transactional serializability and ii) those that provide eventual consistency. The first category includes Beldi [50], Boki [27], and T-Statefun [12]. Beldi implements linked distributed atomic affinity logging on DynamoDB to guarantee serializable transactions among AWS Lambda functions with a variant of the two-phase commit protocol. Boki extends Beldi by adding transaction pipeline improvements regarding the locking mechanism and workflow re-execution. T-Statefun [12] also uses two-phase commit with coordinator functions to support serializability on top of Apache Flink’s Statefun. For eventually consistent transactions, T-Statefun provides an alternative option to two-phase commit using the Sagas pattern. Cloudburst [45] also provides causal consistency guarantees within a DAG workflow. Proposed more recently, Netherite [4] offers exactly-once guarantees and a high-level programming model for Microsoft’s Durable Functions [6], but it does not guarantee transactional serializability across functions.

Serverless Runtimes. Serverless runtimes are an orthogonal line of research to transactional SFaaS. They support execution in a sandboxed runtime environment to make (stateless) serverless execution efficient, performant, secure, and developer-friendly. Faasm [43], and Shredder [51] provide low-latency implementations of SFaaS with WebAssembly support. Nightcore [29] is another runtime that utilizes threads in long-running containers, used by Boki [27].

Dataflow Systems. Support for fault-tolerant execution in the cloud with exactly-once guarantees [7, 17] is one of the main drivers behind the wide adoption of modern dataflow systems. However, they lack a generic and developer-friendly programming model with support for transactions and a natural way to program function-to-function calls. Closer to the spirit of Styx are Ciel [40] and Noria [21]. Ciel proposes a language and an engine for distributed fault-tolerant computations based on functions. At

the same time, Noria solves the view maintenance problem via a dataflow architecture that can propagate updates to clients quickly, targeting web-based, read-heavy computations. Yet, none of the two provide a serializable transactional model for workflows of functions like Styx.

Elasticity in Dataflow Systems. A lot of work has been carried out in dynamic reconfiguration [18, 21, 30] and state migration [14, 23, 25] of streaming dataflow systems over the last few years. These advancements are necessary for providing serverless elasticity in the case of state and compute collocation and enable the leveraging of dataflow graphs as an execution model for *serverless* stateful cloud applications, which is a future goal of Styx.

Transactional Protocols. Besides Aria [37] that inspired the protocol we created for Styx §4, two other protocols fit the requirement of no a-priori read/write set knowledge: Starry [52] and Lotus [53]. Starry targets replicated databases with a semi-leader protocol for multi-master transaction processing. At the same time, Lotus [53] focuses on improving the performance of multi-partition workloads using a new methodology called run-to-completion-single-thread (RCST). Styx makes orthogonal contributions to these works and could adopt multiple ideas from them in the future.

Benchmarks. Due to the absence of benchmarks for (transactional) SFaaS, we used an extended variant of YCSB [11] inspired by T-Statefun’s [12] evaluation. Another benchmark we employ is a variant of Deathstar [19], adapted to SFaaS workloads by [50]. However, it contains a small percentage of transactions (0.5%). Finally, a transactional benchmark that targets OLTP systems is TPC-C [36], and to use it in our evaluation, we had to rewrite it for the SFaaS paradigm.

Summary. To the best of our knowledge, Styx is the first SFaaS system that i) performs deterministic transactions, ii) leverages a dataflow engine that encapsulates both state and function execution, and iii) provides a high-level SFaaS programming model that hides low-level transaction processing primitives from programmers. Finally, Styx outperforms the SotA by more than an order of magnitude in terms of throughput.

10 CONCLUSION

This paper presented Styx, a distributed streaming dataflow system that supports multi-partition transactions with serializable isolation guarantees through a high-level, standard Python programming model that obviates transaction failure management, such as retries and rollbacks. Styx follows the deterministic database paradigm while implementing a streaming dataflow execution model with exactly-once processing guarantees. Styx outperforms the state-of-the-art by at least one order of magnitude in all tested workloads regarding throughput.

ACKNOWLEDGMENTS

This publication is part of project number 19708 of the Vidi research program, which is partly financed by the Dutch Research Council (NWO).

REFERENCES

- [1] [n.d.]. Akka.io. <https://akka.io/>.
- [2] Daniel J. Abadi and Jose M. Faleiro. 2018. An Overview of Deterministic Database Systems. In *Commun. ACM*.
- [3] Michael Armbrust, Tathagata Das, Joseph Torres, Burak Yavuz, Shixiong Zhu, Reynold Xin, Ali Ghodsi, Ion Stoica, and Matei Zaharia. 2018. Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark. In *SIGMOD*.
- [4] Sebastian Burckhardt, Badrish Chandramouli, Chris Gillum, David Justo, Konstantinos Kallas, Connor McMahon, Christopher S Meiklejohn, and Xiangfeng Zhu. 2022. Netherite: Efficient execution of serverless workflows. *Proceedings of the VLDB Endowment* 15, 8 (2022), 1591–1604.
- [5] Sebastian Burckhardt, Chris Gillum, David Justo, Konstantinos Kallas, Connor McMahon, and Christopher S Meiklejohn. 2021. Durable functions: Semantics for stateful serverless. *Proceedings of the ACM on Programming Languages* 5, OOPSLA (2021), 1–27.
- [6] Sebastian Burckhardt, Chris Gillum, David Justo, Konstantinos Kallas, Connor McMahon, and Christopher S. Meiklejohn. 2021. Durable functions: semantics for stateful serverless. *Proc. ACM Program. Lang.* 5, OOPSLA (2021), 1–27. <https://doi.org/10.1145/3485510>
- [7] Paris Carbone, Stephan Ewen, Gyula Fóra, Seif Haridi, Stefan Richter, and Kostas Tzoumas. 2017. State Management in Apache Flink&Reg: Consistent Stateful Distributed Stream Processing. In *VLDB*.
- [8] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. 2015. Apache flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 36, 4 (2015).
- [9] K Mani Chandu and Leslie Lamport. 1985. Distributed snapshots: determining global states of distributed systems. In *TOCS*.
- [10] Cheng Chaoyi, Han, Han Mingzhe, Nuo Xu, Spyros Blanas, Michael D. Bond, and Yang Wang. 2023. Developer’s Responsibility or Database’s Responsibility? Rethinking Concurrency Control in Databases. In *CIDR*.
- [11] Brian F Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. 2010. Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on Cloud computing*, 143–154.
- [12] Martijn de Heus, Kyriakos Psarakis, Marios Fragkoulis, and Asterios Katsifodimos. 2021. Distributed transactions on serverless stateful functions. In *Proceedings of the 15th ACM International Conference on Distributed and Event-based Systems*, 31–42.
- [13] Martijn de Heus, Kyriakos Psarakis, Marios Fragkoulis, and Asterios Katsifodimos. 2021. Distributed transactions on serverless stateful functions. In *DEBS*.
- [14] Bonaventura Del Monte, Steffen Zeuch, Tilmann Rabl, and Volker Markl. 2020. Rhino: Efficient management of very large distributed state for stream processing engines. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2471–2486.
- [15] Akon Dey, Alan Fekete, Raghunath Nambiar, and Uwe Röhm. 2014. YCSB+ T: Benchmarking web-scale transactional databases. In *2014 IEEE 30th International Conference on Data Engineering Workshops*. IEEE, 223–230.
- [16] E. N. Mootaz, Elnozahy, Lorenzo Alvisi, Yi-Min Wang, and David B. Johnson. 2002. A Survey of Rollback-recovery Protocols in Message-passing Systems. *ACM Comput. Surv.* 34, 3 (2002), 34. <https://doi.org/10.1145/568522.568525>
- [17] Raul Castro Fernandez, Matteo Migliavacca, Evangelia Kalyvianaki, and Peter Pietzuch. 2014. Making state explicit for imperative big data processing. In *2014 USENIX annual technical conference (USENIX ATC 14)*, 49–60.
- [18] Avriella Floratou, Ashvin Agrawal, Bill Graham, Sriram Rao, and Karthik Ramasamy. 2017. Dhalion: self-regulating stream processing in heron. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1825–1836.
- [19] Yu Gan, Yanqi Zhang, Dailun Cheng, Ankitha Shetty, Priyal Rathi, Nayan Katarki, Ariana Bruno, Justin Hu, Brian Ritchken, Brendon Jackson, et al. 2019. An open-source benchmark suite for microservices and their hardware-software implications for cloud & edge systems. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 3–18.
- [20] Can Gencer, Marko Topolnik, Viliam Durina, Emin Demirci, Ensar B. Kahveci, Ali Gürbüz Ondrej Lukás, József Bartók, Grzegorz Gierlach, Frantisek Hartman, Ufuk Yilmaz, Mehmet Dogan, Mohamed Mandouh, Marios Fragkoulis, and Asterios Katsifodimos. 2021. Hazelcast Jet: Low-latency Stream Processing at the 99.99th Percentile. In *VLDB*.
- [21] Jon Gjengset, Malte Schwarzkopf, Jonathan Behrens, Lara Timbó Araújo, Martin Ek, Eddie Kohler, M Frans Kaashoek, and Robert Morris. 2018. Noria: dynamic, partially-stateful data-flow for high-performance web applications. In *OSDI*.
- [22] J GRAY. 1978. Notes in Database Operating Systems. *Lecture Notes in Computer Science* 60 (1978), 393–481.
- [23] Rong Gu, Han Yin, Weichang Zhong, Chunfeng Yuan, and Yihua Huang. 2022. Mecas: Latency-efficient Rescaling via Prioritized State Migration for Stateful Distributed Stream Processing Systems. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, 539–556.
- [24] Bas P Harenslak and Julian de Ruiter. 2021. *Data Pipelines with Apache Airflow*. Simon and Schuster.
- [25] Moritz Hoffmann, Andrea Lattuada, Frank McSherry, Vasiliki Kalavri, John Liagouris, and Timothy Roscoe. 2019. Megaphone: Latency-conscious state migration for distributed streaming dataflows. *Proceedings of the VLDB Endowment* 12, 9 (2019), 1002–1015.
- [26] Gabriela Jacques-Silva, Fang Zheng, Daniel Debrunner, Kun-Lung Wu, Victor Dogaru, Eric Johnson, Michael Spicer, and Ahmet Erdem Sariyüce. 2016. Consistent Regions: Guaranteed Tuple Processing in IBM Streams. *Proc. VLDB Endow.* 9, 13, 1341–1352. <https://doi.org/10.14778/3007263.3007272>
- [27] Zhipeng Jia and Emmett Witchel. 2021. Boki: Stateful serverless computing with shared logs. In *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, 691–707.
- [28] Zhipeng Jia and Emmett Witchel. 2021. Nightcore: efficient and scalable serverless computing for latency-sensitive, interactive microservices. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 152–166.
- [29] Zhipeng Jia and Emmett Witchel. 2021. Nightcore: efficient and scalable serverless computing for latency-sensitive, interactive microservices. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 152–166.
- [30] Vasiliki Kalavri, John Liagouris, Moritz Hoffmann, Desislava Dimitrova, Matthew Forshaw, and Timothy Roscoe. 2018. Three steps is all you need: fast, accurate, automatic scaling decisions for distributed streaming dataflows. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, 783–798.
- [31] Tom Killalea. 2016. The Hidden Dividends of Microservices. *ACM Queue* (2016).
- [32] Rodrigo Laigner, Yongluan Zhou, Marcos Antonio Vaz Salles, Yijian Liu, and Marcos Kalinowski. 2021. Data Management in Microservices: State of the Practice, Challenges, and Research Directions. *PVLDB* 14, 13 (2021).
- [33] Leslie Lamport. 2001. Paxos made simple. *ACM SIGACT News (Distributed Computing Column)* 32, 4 (Whole Number 121, December 2001) (2001), 51–58.
- [34] Leslie Lamport, Robert Shostak, and Marshall Pease. 1982. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems* 4, 3 (1982), 382–401.
- [35] Andrea Lattuada, Frank McSherry, and Zaheer Chothia. 2016. Faucet: a user-level, modular technique for flow control in dataflow engines. In *ACM SIGMOD BeyondMR Workshop*. ACM.
- [36] Scott T Leutenegger and Daniel Dias. 1993. A modeling study of the TPC-C benchmark. *ACM Sigmod Record* 22, 2 (1993), 22–31.
- [37] Yi Lu, Xiangyao Yu, Lei Cao, and Samuel Madden. 2020. Aria: a fast and practical deterministic OLTP database. In *VLDB*.
- [38] Yanhua Mao, Flavio P. Junqueira, and Keith Marzullo. 2008. Mencius: building efficient replicated state machines for WANs. In *8th USENIX Symposium on Operating Systems Design and Implementation (OSDI 08)*.
- [39] Derek G Murray, Frank McSherry, Rebecca Isaacs, Michael Isard, Paul Barham, and Martin Abadi. 2013. Naiad: a timely dataflow system. In *ACM SOSP*.
- [40] Derek G Murray, Malte Schwarzkopf, Christopher Smowton, Steven Smith, Anil Madhavapeddy, and Steven Hand. 2011. {CIEL}: A Universal Execution Engine for Distributed {Data-Flow} Computing. In *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*.
- [41] Shadi A. Nohabi, Kartik Paramasivam, Yi Pan, Navina Ramesh, Jon Bringham, Indranil Gupta, and Roy H. Campbell. 2017. Samza: Stateful Scalable Stream Processing at LinkedIn. *Proc. VLDB Endow.* 10, 12, 1634–1645. <https://doi.org/10.14778/3137765.3137770>
- [42] Diego Ongaro and John Ousterhout. 2014. In search of an understandable consensus algorithm. In *ATC*.
- [43] Simon Shillaker and Peter Pietzuch. 2020. Faasm: Lightweight isolation for efficient stateful serverless computing. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, 419–433.
- [44] Pedro Silvestre, Marios Fragkoulis, Diomidis Spinellis, and Asterios Katsifodimos. 2021. Clonos: Consistent Causal Recovery for Highly-Available Streaming Dataflows. In *SIGMOD*.
- [45] Vikram Sreekanti, Chenggang Wu, Xiayue Charles Lin, Johann Schleier-Smith, Jose M Faleiro, Joseph E Gonzalez, Joseph M Hellerstein, and Alexey Tumanov. 2020. Cloudburst: Stateful functions-as-a-service. *arXiv preprint arXiv:2001.04592* (2020).
- [46] Alexander Thomson and Daniel J. Abadi. 2010. The Case for Determinism in Database Systems. In *VLDB*.
- [47] Alexander Thomson, Thaddeus Diamond, Shu-Chun Weng, Kun Ren, Philip Shao, and Daniel J Abadi. 2012. Calvin: fast distributed transactions for partitioned database systems. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 1–12.
- [48] Stephanie Wang, Eric Liang, Edward Oakes, Ben Hindman, Frank Sifei Luan, Audrey Cheng, and Ion Stoica. 2021. Ownership: A Distributed Futures System for Fine-Grained Tasks. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, 671–686. <https://www.usenix.org/conference/nsdi21/presentation/cheng>

- [49] Andreas Wittig and Michael Wittig. 2023. *Amazon Web Services in Action: An in-depth guide to AWS*. Simon and Schuster.
- [50] Haoran Zhang, Adney Cardoza, Peter Baile Chen, Sebastian Angel, and Vincent Liu. 2020. Fault-tolerant and transactional stateful serverless workflows. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*. 1187–1204.
- [51] Tian Zhang, Dong Xie, Feifei Li, and Ryan Stutsman. 2019. Narrowing the gap between serverless and its state with storage functions. In *Proceedings of the ACM Symposium on Cloud Computing*. 1–12.
- [52] Zihao Zhang, Huiqi Hu, Xuan Zhou, and Jiang Wang. 2022. Starry: multi-master transaction processing on semi-leader architecture. *Proceedings of the VLDB Endowment* 16, 1 (2022), 77–89.
- [53] Xinjing Zhou, Xiangyao Yu, Goetz Graefe, and Michael Stonebraker. 2022. Lotus: scalable multi-partition transactions on single-threaded partitioned databases. *Proceedings of the VLDB Endowment* 15, 11 (2022), 2939–2952.